LINKÖPINGS UNIVERSITET

KANDIDATUPPSATS

LIU-IDA/KOGVET-G--11/025--SE

18 HP

Mer lättläst

Påbyggnad av ett automatiskt omskrivningsverktyg till lätt svenska

---

Författare:
Peder Abrahamsson

Handledare:
Arne Jönsson

9 Juni 2011

Sammanfattning

Det svenska språket ska finnas tillgängligt för alla som bor och verkar i Sverige. Därför är det viktigt att det finns lättlästa alternativ för dem som har svårighet att läsa svensk text. Detta arbete bygger vidare på att visa att det är möjligt att skapa ett automatiskt omskrivningsprogram som gör texter mer lättlästa. Till grund för arbetet ligger CogFLUX som är ett verktyg för automatisk omskrivning till lätt svenska. CogFLUX innehåller funktioner för att syntaktiskt skriva om texter till mer lättläst svenska. Omskrivningarna görs med hjälp av omskrivningsregler framtagna i ett tidigare projekt. I detta arbete implementeras ytterligare omskrivningsregler och även en ny modul för hantering av synonymer. Med dessa nya regler och modulen ska arbetet undersöka om det är det är möjligt att skapa system som ger en mer lättläst text enligt etablerade läsbarhetsmått som LIX, OVIX och Nominalkvot. Omskrivningsreglerna och synonymhanteraren testas på tre olika texter med en total lägnd på ungefär hundra tusen ord. Arbetet visar att det går att sänka både LIX-värdet och Nominalkvoten signifikant med hjälp av omskrivningsregler och synonymhanterare. Arbetet visar även att det finns fler saker kvar att göra för att framställa ett riktigt bra program för automatisk omskrivning till lätt svenska.

Innehållsförteckning

Inledning

Enligt Språklagen (2009:600) är det gemensamma språket i Sverige svenska. Alla som är bosatta i Sverige har rätt till att använda och lära sig språket. Svenskan ska gå att använda i alla samhällsområden. Det är även viktigt att bibehålla svenska och dess mångfald (Språklag 2009). På grund av detta är det viktigt att språket anpassas efter läsaren. Svenskan ska vara lätt att förstå både muntligt och skriftligt (Regeringskansliet 2009). Enligt Centrum för lättläst är 25 % av Sveriges vuxna befolkning så pass dåliga på att läsa att de inte uppnår förväntningarna på läskunnighet efter grundskolan (Centrum för lättläst 2002).

Syftet med kandidatuppsatsen var att skapa och implementera fler regler och en synonymhantering i programmet CogFlux för automatisk omskrivning till mer lättläst svenska samt att utvärdera resultatet. Bedömningen av om omskrivningarna ger mer lättläst svenska sker genom att se om texterna som produceras ger ett bättre läsbarhetsvärde på olika mått för läsbarhet, LIX, OVIX och nominalkvot. Ett ökat läsbarhetsvärde bör inte påverka textens innehåll så att för mycket information faller bort. Dessa kriterier kommer att underlätta läsförståelsen och är därmed enligt språklagen (2009:600) positivt för det svenska språket, då fler individer kan ta del av det Svenska språket.

Kandidatuppsatsen är en del i ett större projekt kallat EasyReader. Syftet med huvudprojektet är att göra svenska språket tillgängligt för personer som av olika anledningar inte har samma förutsättningar att ta in informationen från text. För att underlätta läsningen för denna grupp människor så skrivs texter om för att bli mer lättlästa. Att manuellt skriva om texter är tid och resurskrävande. Därför finns det önskemål om att kunna skriva om texter automatiskt till att bli mer lättlästa.

Arbetet utgår från ett tidigare arbete kallat CogFLUX som är ett system för automatisk omskrivning till lättläst svenska. CogFLUX är skapat på ett vis som gör att de lämpar sig bra för att bygga på med fler omskrivningsregler (Rybing & Smith 2009).

1.1 Avgränsningar

Eftersom kandidatarbetet bygger på ett program som arbetar med svensk text kommer arbetet fokusera på och endast bearbeta svenska texter. Kandidatarbetet kommer inte att avgränsas till att endast arbeta med en viss typ av texter. Att endast inrikta sig på en typ av text skulle innebära att programmet får en spetskompetens för just den sortens text. Kandidatarbetet upplevs mer intressant och blir bredare då det kan förenkla olika sorters texter. Med ett program som kan förenkla flera olika typer av texter ges en god grund att diskutera kring huruvida de olika texterna blir mer lättlästa än andra. Många läsare skulle dra nytta av mer lättlästa versioner av vissa typer av texter. Dock är det problematiskt att utveckla denna spetskompetens, då få undersökningar utförts kring vad som gör den specifika texten mer lättläst än andra texter inom samma område.

Det kommer inte genomföras någon analys med hjälp av användartester, det vill säga testpersoner. Analysen kommer endast att baseras på mått för läsbarhet. Kandidatarbetet kommer inte att resultera i några nya läsbarhetsmått, utan kommer att använda mått som andra redan tagit fram och redan är implementerade i programmet.

Teoribakgrund

I denna del av rapporten diskuteras teori kring lättlästa texter. Mått på läsbarhet presenteras och tidigare arbeten som ligger till grund för detta tas upp.

2.1 Lättlästa texter

När någon pratar om lättlästa texter dras tankarna lätt åt småbarnsböcker med många bilder och en eller två meningar på varje sida. Även om de texterna antagligen skulle hamna inom ramen för vad som är en lättläst text så måste det även gå att hitta lättlästa texter som passar en vuxen målgrupp. Men vad är egentligen en lättläst text? Centrum för lättläst har tagit fram flera olika egenskaper hos lättlästa texter. Dessa egenskaper kan även ses som kriterier för att kunna kalla en text lättläst. Ju fler egenskaper som stämmer överens med texten ju mer lättläst är den. Egenskaperna är steglösa och när man undersöker om egenskapen finns är det svårt att säga att den finns fullt ut eller inte alls. Det vill säga att texten kan stämma mer eller mindre överens med en egenskap (Centrum för lättläst 2002).

Skriv tydligt – En viktig egenskap som gör texter lättlästa är att texten är tydlig. Det ska inte vara nödvändigt att läsa om en mening flera gånger för att kunna förstå vad som står. Många meningar som är svåra att förstå går att förenkla genom att lägga till en förklaring. Är något svårt att förklara är det bättre att ta bort den delen ur texten.

Viktigast först – Det viktigaste i texten bör stå i början. Den första informationen är enklast att ta in eftersom informationen från övriga texten ännu inte är förstådd. Ju mindre information som läsaren behöver ta hänsyn till desto lättare är det att förstå informationen. Därför är det även viktigt att endast skriva det som behöver skrivas. För att hålla texten ännu mer lättläst så bör den skrivas så att det endast är en information per rad.

Undvik omskrivningar – Skriv det som menas utan att använda metaforer eller bildspråk. Läsaren kan tolka texten bokstavligt vilket inte alltid är meningen. I meningen ”Hon är en riktig pudding” kan personen som läser tolka det som att personen faktiskt är en pudding. Då är det istället bättre att skriva ”Hon är riktigt snygg” för att få bort metaforen.

Använd få olika personer och platser – Blandas för många olika personer i texten så blir det svårare att hålla reda på vem som är vem. Samma gäller för antalet olika platser. Tas

för många olika platser upp blir det svårare att hänga med i var något hänt eller händer.

Skriv kronologiskt – Skriv så att det som sker först även står först. Undvik även att använda tidsuttryck. En del läsare kan endast skilja på det som hänt och det som händer.

Använd aktiv form – Texten blir mer lättläst om passiv form undviks. Skriv ”Hästen dricker vatten” istället för ”Vattnet dricks av hästen”.

Undvik svåra ord – Långa ord är svårare att läsa eftersom de består av flera tecken som kan vara svåra att hålla samman till en enhet. Det är bättre att använda vanligare ord än ord som inte är lika utbredda. Vet inte läsaren vad ordet betyder uppstår ett avbrott i läsningen. Det är bättre att byta ut ord mot synonymer om de är svåra att läsa. Är ett ord ovanligt och svårt att förstå bör det förklaras.

Undvik negationer – Flera negationer i samma mening gör den svåröverskådlig och kan behöva läsas igenom igen för att den ska bli begriplig. Det är bättre att skriva om meningen så att den innehåller så få negationer som möjligt. Meningen ”Hon är en av de minst besvärande ointressanta människor, som det varit mig en total brist på nöje, att inte kunna undgå att träffa.” gör det tydligt negationer kan få texten svårbegriplig.

Använd huvudsatser – Bisatser bryter upp meningen så det blir svårare att få grepp om helheten. Det är bättre att dela upp bisatser i flera meningar. Subjunktioner bör också undvikas. Ord så som ”om”, ”när” och ”eftersom” bör inte inleda en mening.

Skriv ut förkortningar – Det är bättre att skriva vad något faktiskt är istället för att använda ett substitut som förkortningar. Det är bättre att skriva ”det vill säga” istället för ”d.v.s. ”. Även pronomen är bättre att skriva ut vem eller vad den syftar till.

Beskriv siffror – Siffror i en text kan vara svåra att förstå. Många saknar förmågan att enkelt kunna relatera till nummer och mått. Relatera numret till något på ett konkret vis så att numret blir begripligt. Avstånd blir också lättera att förhålla sig till om de relaterar till hur lång tid de tar åka istället för att vara beskrivna i mil och kilometer (Freyoff et al 1998, Davidsson et al 2002, Centrum för lättläst 2002, Decker 2003).

2.1.1 Problem med att skriva tydligt

Att skriva tydligt behöver inte bara betyda att begrepp som är enkla att förstå används det innebär också att texten inte ska vara lätt att misstolka och helst inte alls kunna misstolkas. Därför ska tvetydiga begrepp och tvetydiga meningar undvikas. Det går inte alltid att helt få bort tvetydighet i meningar. Ord som tillhör ordklassen prepositioner

kan ofta tolkas på mer än ett vis även om meningen är en mycket enkel sådan. Ett exempel på en sådan mening är ”Mannen tittar på kvinnan med kikaren” där det inte är uppenbart om det är kvinnan eller mannen som har kikaren. Även om meningen skulle skrivas om så att det blir tydligare vem som har kikaren så är det mycket svårt att få meningen att bara kunna tolkas på ett vis. Skulle den till exempel skrivas om till ”Mannen tittar med kikaren på kvinnan” så blir det enklare att förstå att det är mannen som tittar med hjälp av kikaren. Den nya meningen kan även den tolkas på flera vis, befinner sig kikaren på kvinnan eller är kikaren och mannen kompisar som tillsammans tittar på kvinnan? Dessa alternativa tolkningar uppfattas som mycket mindre troliga vilket gör att det inte är lika lätt att misstolka meningen. (Jörgensen & Svensson 2004, Centrum för lättläst 2002)

2.2 Läsbarhetsmått

Ett sätt att bedöma texters lättlästhet är genom olika läsbarhetsmått. Till skillnad från redan nämnda riktlinjer för lättlästa texter ger läsbarhetsmått kvantitativ data i form av värden. Det finns flera sorters läsbarhetsmått. De beskrivs nedan i varsin del. Gemensamt för alla måtten är att de kan användas både på texten som ska göras om samt den nya texten för att direkt se en skillnad. Även om måttet i sig inte visar att texten nu kan klassas som lättläst så går det direkt att se om texten blivit mer lättläst än den var innan. Läsbarhetsmått är på grund av detta lämpligt till utvärdering av hur lättlästa texter är. Värt att minnas är att texterna då endast blir mer lättlästa utifrån just måtten.

2.2.1 Läsbarhetsindex

Läsbarhetsindex från och med nu förkortat till LIX är ett mått på läsbarhet som tar hänsyn till antalet meningar, ord och längden på ord för att räkna ut läsbarheten hos en text (Björnsson 1968). Ju lägre LIX-värdet är desto mer lättläst anses texten vara. LIX-värdet räknas ut med hjälp av följande formel:

LIX = Antal(ord)/Antal(meningar) + Antal(långa ord)/Antal(ord) *100

Ord som innehåller mer än sex tecken definieras som långt. Ett uträknat LIX-värde har inget egentligt egenvärde eftersom det enbart är ett tal. Generellt är att ju lägre LIX-värde desto mer lättläst text. För att kunna jämföra måtten effektivt går det att använda sig av nedanstående tabell (Björnsson 1968).

Tabell 1. Översiktstabell av LIXvärden på olika texter.

Texttyp	LIXvärde
Mycket lättläst, barnböcker	<30
Lättläst, skönlitteratur, populärtidningar	30-40
Medelsvår, normal tidningstext	40-50
Svår, normalt värde för officiella texter	50-60
Mycket svår, byråkratisvenska	> 60

2.2.2 Nominalkvot

Nominalkvot är ett mått på informationstäthet i texten. Ju lägre värde desto mer lättläst text. För att räkna ut nominalkvot används följande formel :

Nominalkvot = Antal(nomen+prepositioner+particip) /Antal(pronomen+adverb+verb) x 100

Normalvärdet för nominalkvot ligger på 100 och kan jämföras med en tidningstext. (Josephson & Melin 1990, Wilhelmsson 2007, Rybing & Smith 2009).

2.2.3 Ordvariationsindex

Ordvariationsindex, från nu kallat OVIX, är ett mått på hur många unika ord som finns i texten, alltså de ord som endast förekommer en gång, jämfört med totalt antal ord. Ju fler unika ord i förhållande till antalet ord totalt desto mer svårläst anses texten vara. Desto lägre OVIX-värde ju färre unika ord och mer lättläst text. Formeln för hur OVIX räknas ur beskrivs nedan:

OVIX = Antal(unika ord)/Alla ord *100

Textens längd påverkar hur OVIX-värdet ska tolkas i och med detta det inte finns några lika bra riktlinjer för vilka värden som kan klassa en text som lättläst. En längre text innehåller flera ord men inte nödvändigtvis fler unika ord (Rybing & Smith 2009).

2.2.4 Hur väl fungerar måtten?

De tre mått som beskrivs för att värdera hur lättlästa texterna är har både brister och fördelar. LIX är det vanligaste svenska måttet för läsbarhet och är därför enkelt att använda som gradering på lättlästhet. Det finns en tabell att jämföra på vilken svårighetsnivå texten befinner sig genom att använda LIX. En brist som LIX, och även de andra läsbarhetsmåtten, har är att måttet inte tar någon hänsyn till ordens svårighetsgrad. Ett långt ord som *fotbollsspelare* är ett vanligt ord och blir därför lättläst även om det försämrar LIX-värdet. Ett kort ord som *avog* är däremot svårare att förstå men ger ett lägre LIX-värde. Ett annat exempel är att LIX skulle anse *spelade* svårare än *spelar* eftersom det innehåller fler tecken. Flera av kriterierna för vad som kännetecknar en lättläst text tar inte måtten någon hänsyn till. Om texten är skriven med många metaforer eller om handlingen hoppar fram och tillbaka i tiden är moment som gör texten mer svårläst men de kan ändå producera utmärkta resultat på läsbarhetsmåtten. Inget av måtten kräver att texten måste vara grammatiskt korrekt. Meningar kan vara ofullständiga och ord kan vara felaktigt böjda utan att det skulle märkas på något av måtten. Även om måtten har brister så ger de en tydlig riktlinje för om en text blir mer lättläst eller inte (Mühlenbock & Johansson 2009, Lundberg & Reichenberg 2008, Centrum för lättläst 2002).

2.3 Deckers omskrivningsregler

För att kunna skapa ett automatiskt omskrivningsprogram för mer lättläst svenska krävs det kunskap om ur texten ska modifieras. Decker (2003) har tagit fram syntaktiska omskrivningsregler för mer lättläst svenska. Dessa regler är framtagna och skrivna på ett sådant vis att de enkelt ska kunna läggas in i ett datorprogram. För att hitta och formulera dessa omskrivningsregler arbetade Decker med texter som var omskrivna till lätt svenska utan användning av ett automatiskt omskrivningsprogram. De texter som användes som underlag för arbetet var hämtade från Invandrartidningen. Invandrartidningen är en tidning som riktar sig till personer med utländsk bakgrund. Tidningen ges ut på flera olika språk och förekommer även som omskrivning till lättläst svenska. Omskrivningen av texterna gjordes på tidningsredaktionen med hjälp av riktlinjer från centrum för lättläst. Dessa texter granskades sedan manuellt för att få fram vad som kännetecknar en lättläst text. Decker jämförde den omskrivna varianten av texten till lättläst svenska med originalet för att hitta skillnader mellan texterna. Texterna ordklasstaggades och parsades så att Decker kunde se deras uppbyggnad. Utifrån de

tendenser Decker kunde finna mellan den lättlästa varianten på texten och den som var skriven på vanlig svenska skapades 467 förenklingspar. Ett förenklingspar består av en fras från originaltexten och dess förenklade motsvarighet från den omskrivna varianten. Utifrån dessa förenklingspar kom Decker fram till 25 omskrivningsregler som förändrar syntaxen i texten till en mer lättläst variant. Ett exempel på en sådan regel som Decker kom fram till är ap(adj1+konj+adj2) → ap(adj2) som betyder att en adjektivfras (ap) som innehåller ett adjektiv (adj1) en konjunktion (konj) och ett till adjektiv (adj2) görs om till en adjektivfras som innehåller bara det ena av adjektiven (Decker 2003). Ett exempel på en mening som skulle skrivas om enligt denna regel är ”Hon genomgår en *svår och dyr behandling*”. Den delen som är skriven kursivt visar vad i meningen som matchar regeln. Den omskrivna varianten av meningen eligt regeln blir ”Hon genomgår en *dyr behandling”*. Alla Deckers regler går att finna i bilaga 1.

2.4 CogFLUX

CogFLUX är ett datorprogram som bygger vidare på Deckers arbete. CogFLUX är skapat för att förenkla svenska texter automatiskt och skapades av Rybing och Smith (2009). CogFLUX förenklar svenska texter och den automatiska omskrivningen kan enbart användas till detta språk. Programmet är bara påbörjat och det finns stort utrymme till förbättring. Rybing och Smith (2009) beskriver det som ”en verktygslåda för vidare utveckling”. Programmet är indelat i tre större delområden som hanterar texter. Det första området granskar och gör om texten så att den blir mer hanterbar för ett datorprogram. I det andra området av programmet sker den omarbetning som gör texten mer lättläst. Här sker den förändring som är syftet med arbetet. I den här delen av programmet finns en modul för hantering av förkortningar samt en modul för omskrivning enligt Deckers regler. Tretton utav Deckers regler finns implementerade i programmet. I den sista delen av programmet tas hjälpmedlen bort så att texten åter blir sammanhängande. I den här delen sker även en utvärdering av den nya texten genom med hjälp av de tre läsbarhetsmåtten LIX, OVIX och Nominalkvot som beskrivits tidigare i rapporten. Läsbarhetsmåtten är redan utarbetade och implementerade i programmet. Innan reglerna som är implementerade i CogFlux kan arbeta om texten behöver texten ordklasstaggas och parsas (Rybing & Smith 2009).

2.4.2 Ordklasstaggning

Ordklasstaggning innebär att varje ord i en text får beskrivet vilken ordklass det tillhör. Ordklasstaggning är nödvändigt för att ett datorprogram ska kunna hantera texten uttifrån vilken ordklass orden i texten tillhör. Det går att arbeta mer generellt med en text som är ordklasstaggad. Med en ordklasstaggad text går det att göra analyser och arbeta med texten beroende på vilka ordklasser orden tillhör och inte enbart med vilka ord och tecken som finns i texten. När en text är ordklasstaggad går det till exempel att räkna alla substantiv som finns i texten utan att behöva ha en lista med ord som är substantiv att jämföra mot. Granska Tagger är en ordklasstaggare som används i CogFLUX vilket beskrivs nedan. Granska Tagger klarar av att korrekt tagga 96,3 % av alla ord som den testades på (Carlberger & Kann 1999).

2.4.3 Parsning

Parsning innebär att man skriver ut hur fraserna i en text är uppbyggda. Genom att parsa en text går det att hantera den efter hur fraserna ser ut och är uppbyggda. För att kunna parsa texten måste en ordklasstaggning ha skett alternativt är ordklasstaggningen inbyggd i parsen. MaltParser är en parser som utvecklats vid Växsjö universitet. MaltParser använder sig av färdiga ordklasstaggade texter för att skapa frasstrukturer. MaltParser är baserat på Support Vector Machines som klassificerar data. MaltParser ger en F1 poäng mellan 75 och 80 vilket räknas som en relativt god prestanda. F1 är ett mått på exakthet utifrån korrekta parsningar och antalet totala parsningar (Hall 2006, Nivre & Hall 2009, Rybing & Smith 2009).

2.5 Folkets synonymlexikon Synlex

Förändring av texter som gör dem mer lättlästa behöver inte enbart ske genom syntaktiska omskrivningar, det kan göras genom att byta ut ord. Synlex är ett synonymlexikon framtaget med hjälp av den svenska webbplatsen Lexin on-line, som är ett uppslagsverk, där användarna fick bedöma huruvida de ansåg att ord var synonymer med varandra (Kann 2003). Synonymer är ord som betyder samma sak. Det är mycket sällan som synonymerna är helt utbytbara med varandra, utan det förekommer oftast skillnader i tillexempel stilvärde (Kann 2004). Användarna på Lexin on-line fick gradera synonymerna på en skala från 0 (instämmer inte) till 5 (instämmer fullt). Synonymer

som slutgiltigt blev graderad som tre eller högre används i Synlex. Synonymerna bedömdes inte enbart av användarna utan de justerades även med hjälp av andra metoder. Bland annat togs olika värden fram för att bedöma hur ofta synonymerna förekommer i liknande sammanhang (Kann 2003). Utdrag ur folkets synonymlexikon Synlex visas i bilaga 2.

Metod

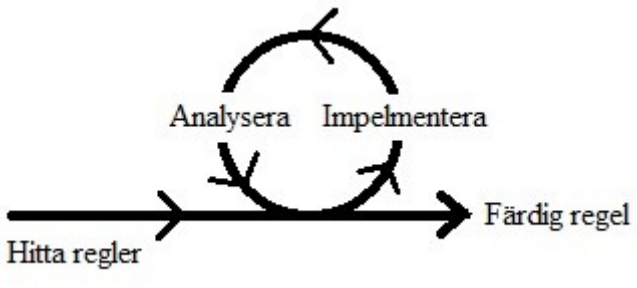
Metoden som användes bestod i huvudsak av tre delar. Dessa delar i metoden var att ta reda på vad för omskrivningar av en text som kan göras för att få den mer lättläst, implementera den typen av omskrivningar i programmet och slutligen analysera resultatet av den nya texten jämfört med den gamla. Reglerna granskades i huvudsak genom att läsa igenom meningar som är förenklade och se efter så att språket är bra. Om meningen som var omskrivna inte blev sammanhängande eller tydliga redigerades regeln så att ett bättre resultat uppnåddes. För att förtydliga arbetets gång pressenteras denna i bild 1.

För att ta reda på hur en text kan göras mer lättläst används i huvudsak omskrivningsregler som Decker (2003) tagit fram utifrån syftet att ta fram regler som ett datorprogram kan hantera. Vilka regler som implementeras i programmet avgörs utifrån hur gott stöd programmet har för dem.

Implementationen av regler sker i programfilen med namnet *deckerrules.smec*. I denna programfil kan omskrivningsregler göras enligt formatet Frastyp-Ordklassföljd → Frastyp- Ny ordklassföljd (Förälderfras). Den första delen, Frastyp-Ordklassföljd, beskriver hur det som ska förenklas ser ut i originaltexten och den andra delen, Frastyp-Ny ordklassföljd (Förälderfras), beskriver hur det kommer se ut efter omskrivningen och förenkling.

Resultatet av omskrivningen ges direkt av programmet i form av en ny version av texten samt resultatet av LIX, OVIX och nominalkvot. Är resultatet tillfredsställande sparas den regel som användes och nästa regel påbörjas. Om önskat resultat inte uppnås modifieras regeln så att den fungerar. Ett önskat resultat innebär att texten har omskrivits på så vis som regeln sagt att den skulle.

Bild 1. Översiktsbild över arbetets gång.



Resultat

Resultatet inkluderar både implementeringarna, synonymhanteraren och analysen. Analysen utvärderar reglerna och synonymhanteraren utifrån läsbarhetsmåtten, LIX, OVIX och nominalkvot.

5.1 Implementeringar

Detta avsnitt beskriver de nya regler som implementerats i CogFLUX. Alla regler baseras på de regler som Decker tagit fram genom sitt arbete. Det finns vissa skillnader i hur reglerna skrivs jämfört med hur Decker beskriver dem och därför kommer de regler som implementerats i det här arbetet att nämnas som Syntaktisk omskrivningsregel eller SOR. Alla SOR som implementerats tar bort delar av en fras. De representeras här på samma vis som de är skrivna i programkoden. REPL// betyder replace eller byt ut och beskriver vilken operation som görs på fraserna. Därefter beskrivs vilken fras omskrivningen gäller och vilken struktur frasen har. Det som kommer efter → är hur frasen kommer att se ut efter omskrivningen. Sist i SOR innanför parenteser står det om frasen måste ha en specifik föräldrafras för att kunna matchas mot texten. Behövs ingen föräldrafras så placeras istället en brädgård på den platsen. Innan den eventuella föräldrafrasen står det Â§P. Dessa tecken ändras aldrig och har ingen operativ betydelse utan markerar bara vart en fras slutar. Alla SOR är testade på de exempelmeningar som anges men kan skriva om alla meningar som har samma frasstruktur som exempelmeningarna.

Den första SOR som implementerats är:
REPL//NP-DT NN S → NP-NN S Â§P(PP)
Den här SOR ska ta bort determinatorer (DT) som föregår ett substantiv (NN) följt av en bisats (S) i en nominalfras (NP-). För att determinatorn ska tas bort ur nominalfrasen krävs det även att nominalfrasen i sin tur finns inom en prepositionsfras (PP). Prepositionsfrasen är då en förälder till nominalfrasen. Denna SOR skriver om meningar som ser ut som följande ”Dom är till för *de* elever som redan går på gymnasiet” till ”Dom är till för elever som redan går på gymnasiet”. Det krävs att frasen har en prepositionsfras som förälder för att undvika att *en* tas bort från meningar som till exempel ”Det är en stol som är brun.”.

Den andra SOR som implementerats är:
REPL//AP-AP KN AP -> AP-AP Â§P(#)
Den här SOR tar bort en adjektivfras (AP) och en konjunktion (KN) från en adjektivfras (AP-) som innehåller två adjektivfraser och en konjunktion. Regeln skriver om meningar som ser ut på detta vis: ”Han har en stor *och grön* tröja.” till ”Han har en stor tröja.”. Denna SOR har inget krav på någon föräldrafras eftersom alla förekomster av ett adjektiv följt av ett och följt av ett till adjektiv ska matchas.

Nästa SOR ska ta bort beskrivningar om vart saker kommer ifrån. Regeln ser ut som följer:
REPL//PP-PP NP -> Â§P(S)
Regeln betyder ta bort en prepositionsfras (PP-) som innehåller en prepositionsfras (PP) följt av en nominalfras (NP). Det den säger är egentligen hitta den här frasen och byt ut den mot en ingenting, men i praktiken blir samma sak som att ta bort. Denna SOR gör om meningen ”Det är importerade ostron från island.” till ”Det är importerade ostron.”.

De två följande SOR är mycket lika varandra. Båda är till för att blocka bort bisatser som fungerar som adjektiv. Reglerna ser ut som följer
REPL//NP-NN MID S MID -> NP-PM Â§P(S)
och
REPL//NP-PM MID S MID -> NP-PM Â§P(S)
Regeln säger ta bort en bisats (S) som ligger mellan två kommatecken (MID) efter ett substantiv (NN/PM) i en nominalfras (NP-).
Dessa SOR skriver om meningar som ser ut som följer: ”Stolen, som är grön, kan inte röra sig själv.” Och gör om dem till ”Stolen kan inte röra sig själv.”. Det första ordet *stolen* kan bytas ut mot ett namn för att matcha den andra regeln av de båda reglerna.

5.2 Synonymhantering

För att på ett effektivt sätt hantera ord som bör bytas ut mot andra lättare ord skapades en ny klass i programmet, vilken kallades SynonymsModule. Den nya synonymhanteringsklassen är till för att byta ut svåra ord mot en synonym som är mer lättläst. Det första SynonymsModule gör är att hämta texten som ska modifieras. Den använder samma text som används för att hantera SOR det vill säga en text som är

ordklasstaggad och parsad. I den textfilen är varje ord redan uppdelat i enskilda enheter vilket annars skulle behövas för att smidigt kunna byta ut ord mot andra ord. Sen läser modulen in synonymordlistan och gör om den så att den blir enkel för programmet att hantera. Endast ord som har synonymer som är kortare än originalordet byts ut. Varje ord kollas mot synonymordlistan en gång. Hittas ett ord som har en eller flera synonymer så byts det ordet ut mot den synonym som har minst antal tecken. Ett exempel på ett ord som byts ut är abdikera som byts ut mot avgå.

5.3 Analysering

Alla regler och Synonymhanteraren har testats på tre olika texttyper för att få ut läsbarhetsvärden efter omskrivningarna. Tabellerna visar hur värdena förändrats för en viss texttyp. Tre olika körningar har gjorts för varje text. En körning där endast synonymhanteraren har använts, en körning där endast de nya SOR använts och en körning där både synonymhantering och SOR använts. Även tre körningar har gjorts där alla tre texttyperna har använts som indata. De tre texterna som använts är en text från en tidningsartikel, en text från en skönlitterär bok och en formell text som kommer från försäkringskassan. Tidningsartikeln och den formella texten innehåller ungefär tjugofem tusen ord. Den skönlitterära texten innehåller ungefär femtiotusen ord.

I körningen där tidningsartikeln använts som indata blev resultatet ett något förbättrat LIX-värde både för körningar med SOR och för körningar med synonymhanteraren. Används både synonymhanterare och SOR blir värdet lite bättre än om de används var för sig. OVIX-värdet är oförändrat i körningen med synonymhanteraren och något försämrat vid användningen av SOR. Används båda blir OVIX-värdet också något sämre. Nominalkvoten är oförändrad när endast synonymhanteraren används men ganska mycket lägre vid användning av SOR. Används båda ges också ett tydligt sänkt värde på nominalkvoten.

Tabell 2. Översiktstabell av läsbarhetsmåttens förändringar av en tidningsartikel

Tidningsartikel				
Läsbarhets-mått	Värden före körning	Synonymhanterare	SOR	Synonymhanterare och SOR
LIX	55	52	53	49
OVIX	22	22	24	24
Nominalkvot	126	128	85	88

Körningar med en skönlitterär text ger liknande resultat som för körningarna med tidningsartikeln. Synonymhanteraren ger ett något lägre LIX-värde och något högre nominalkvot. Synonymhanteraren gav en knappt märkbar förbättring i OVIX-värdet. SOR gav en liten förbättring på LIX-värdet och en klar förbättring på nominalkvoten. Vid körningar med både SOR och synonymhantering blev LIX-värdet mer förbättrat, OVIX oförändrat och nominalkvoten förbättrad men något högre än vid körningar med endast SOR.

Tabell 3. Översiktstabell av läsbarhetsmåttens förändringar av en skönlitterär text

Skönlitterär text				
Läsbarhets-mått	Värden före körning	Synonymhanterare	SOR	Synonymhanterare och SOR
LIX	44	40	42	38
OVIX	14	13	14	14
Nominalkvot	66	68	41	43

Även i den formella texten syns samma mönster. Synonymhanteraren ger ett förbättrat LIX-värde och ett nästan oförändrat värde på OVIX och Nominalkvot. SOR ger även de ett bättre LIX-värde och ett något försämrat OVIX-värde. Både SOR och synonymhanteraren tillsammans ger ett bättre LIX-värde än någon av dem ensamma, men ger också ett högre OVIX-värde än vad någon av dem ensamma ger. Nominalkvoten blir nästan lika mycket förbättrad som vid körningar med bara SOR och mycket bättre än vid körning med bara synonymhanteraren.

Tabell 4. Översiktstabell av läsbarhetsmåttens förändringar av en formell text

Formell text				
Läsbarhets-mått	Värden före körning	Synonymhanterare	SOR	Synonymhanterare och SOR
LIX	51	46	49	44
OVIX	6	7	7	8
Nominalkvot	122	126	91	92

Reglerna ger fortfarande ett förbättrat LIX-värde och nominalkvotvärde. Synonymhanteraren ger ett bättre LIX-värde, OVIX är oförändrat och nominalkvoten något förändrad. När både SOR och synonymhantering används blir LIX-värdet bättre än om endast synonymhanteraren används, detta trots att körningen med enbart SOR gav ett försämrat LIX-värde.

Tabell 5. Översiktstabell av läsbarhetsmåttens förändringar av en tidningsartikel, skönlitterär och formell text då dessa tillsammans används som indata.

Tidningsartikel, skönlitterär text & formell text				
Läsbarhets-mått	Värden före körning	Synonymhanterare	SOR	Synonymhanterare och SOR
LIX	49	45	46	42
OVIX	11	11	12	11
Nominalkvot	90	92	61	62

Från dessa värden går det att säga att synonymhanteraren kan förbättra LIX-värden men inte göra någon direkt skillnad på varken OVIX eller Nominalkvot. SOR kan också förbättre LIX-värden och tillsammans med synonymhanteraren ge ett tydligt bättre LIX-värde. SOR är även mycket bra på att sänka nominalkvoten.

Resultaten från körningarna signifikanstestades med hjälp av ett beroende t-test (Heiman 2001). SOR sänker både LIX och nominalkvot signifikant med ett konfidensintervall på 99%. SOR gav ingen signifikant skillnad på OVIX. Synonymhanteraren sänkte LIX signifikant med ett konfidensintervall på 99%. Synonymhanteraren visade ingen signifikant skillnad på OVIX. Nominalkvoten ökade signifikant med konfidensintervall på 95% när synonymhanteraren analyserades. Då både Synonymhanteraren och SOR användes gav det signifikant lägre LIX med ett konfidensintervall på 99,9%. OVIX

visade inte på någon signifikant skillnad då både SOR och synonymhanterare använts. Nominalkvoten sänktes signifikant med ett konfidensintervall på 99% i körningarna med både synonymhanteraren och SOR.

Alla SOR testades var för sig på tidningsartikeln för att se om någon regel stod för mer av ändringarna på läsbarhetsmåtten.

Tabell 6. Översikt körningar med varje regel var för sig.

SOR	LIX	OVIX	Nominalkvot
Tidningsartikel utan regler	55	22	126
REPL//NP-DT NN S → NP-NN S Â§P(PP)	56	22	126
REPL//AP-AP KN AP → AP-AP Â§P(#)	55	22	126
REPL//PP-PP NP → Â§P(S)	53	24	85
REPL//NP-NN MID S MID → NP-NN Â§P(S)	55	22	126
REPL//NP-PM MID S MID → NP-PM Â§P(S)	55	22	126

Värdena från dessa körningar visar tydligt vilken regel som påverkar mest REPL//PP-PP NP → Â§P(S). Det var även den regel som hade överlägset flest matchningar på texten generellt. De andra reglerna hade bara ett fåtal matchningar. För att få ett mer rättvist resultat över reglernas respektive värde kan det krävas att antalet matchningar tas in i beräkningen. Görs ingen förändring av texten går det heller inte att så någon förändring i läsbarhetsmåtten.

Diskussion

I detta kapitel tas brister i programmet upp samt förslag på hur dessa kan förbättras. Även nya funktioner inför fortsatt arbete tas upp.

6.1 Mått på informationsinnehåll

För att ta reda på om det har försvunnit mycket information från texten skulle det behövas något mått på informationsinnehåll. Alla omskrivningsregler som finns implementerade i programmet tar bort något från texten. Det är inte önskvärt att ta bort information som har en betydande roll för textens innehåll. Det går att räkna antalet värdeord (substantiv, adjektiv och verb) före och efter att texten förenklats. Innehåller den förenklade varianten ungefär lika många värdeord så bör informationsinnehållet vara nära originaltexten. Om istället antalet värdeord är långt färre i den nya texten kan det tolkas som att information har försvunnit på vägen. Av värdeorden är det endast adjektiv som påverkas av reglerna i och med att de tas bort vilket skulle kunna ses som att information försvunnit.

6.2 Förbättring av SOR

Även om de regler som implementerats i CogFLUX förändrar specifika meningar på ett önskvärt vis så ger de även förändringar på meningar som inte ska ändras. För att kunna undvika detta krävs generellt att reglerna blir mer specifika så att de endast ger en matchning på den sortens text som ska ändras. Även om Decker kunde hitta och beskriva 25 stycken olika regler för syntaktiska omskrivningar så betyder inte det att varje förekomst av den frasen bör skrivas om enligt regeln. Ett exempel på en sådan regel är ap(adj1+konj+adj2) → ap(adj2) som är inskriven i programmet som REPL//AP-AP KN AP → AP-AP Â§P(#). Denna regel matcher många fler fall än de som bör ändras, vilket beskrivs senare i detta kapitel.

En regel som enligt Decker (2003) underlättar läsbarheten är implementerad i CogFLUX: REPL//AP-AP KN AP → AP-AP Â§P(#)
Denna regel tar bort ett adjektiv av två adjektiv i samma mening och konjunktionen som binder de samman, till exempel: *Stor och fin hatt → Stor hatt.*
Den här regeln är inte helt optimalt implementerad. Det skulle önskas att den tar bort den första förekomsten av ett adjektiv samt konjunktionen men som den ser ut nu tar den bort det senare adjektivet i frasen. För att förbättra regeln krävs det någon form av

markering som talar om vilken adjektivfras som ska vara kvar. Ett sätt att förbättra regeln är att skriva om programmet så att ett nummer kan skickas med den fras som ska behållas. Exempel REPL//AP-AP KN AP → AP-AP(2) Â§P(#) där nummer 2 i parentesen efter adjektivfrasen skulle syfta på den senare förekomsten av en adjektivfras. Skulle regeln istället skrivas på detta vis REPL//AP-AP KN AP → AP-AP(1) så syftar AP(1) till den första förekomsten av en adjektivfras i ursprungsfrasen.
Att kunna lägga på en markering för exakt vilka delar i en omskrivning som önskas vara kvar skulle förbättra möjligheten till fler automatiska omskrivningar som med den nuvarande programvaran inte kan hantera, nämligen omflyttningar av ordföljden. Som det ser ut nu går det att plocka bort delar ur texten men det går inte att flytta ordningen på orden. Ett annat problem med den omskrivningsregeln är att den matchar alla konjunktioner även om det endast är vid tillfällen där ordet ”och” förekommer som en ändring ska ske. Ett exempel på en fras som regeln matchar är ”Bilen var grön eller blå” vilket skrivs om av programmet till ”Bilen var grön”. För att kunna hantera den här typen av problem krävs möjligheten att skriva till kriterier vilka använder de faktiska orden i texten. Antingen skulle det gå att lägga till möjligheten att byta ut en ordklass mot ett specifikt ord så att regeln istället skulle se ut något i stil med REPL//AP-AP #T och AP → AP-AP. Att skriva regler på det viset ger friheten att använda orden istället för ordklasserna som riktlinje till vilka fraser som kan matchas av reglerna. Tecknet # med ett direkt efterföljande T är bara ett exempel på en markering för att programet ska förstå att det nästa teckensekvensen ska behandlas som en text och inte som en ordklass eller fras. Ett annat exempel på ett vis som skulle kunna fungera på är att lägga till en lista direkt efter ordklassen med ord som inte ska matchas eller ord som ska matchas. Regeln skulle då kunna se ut så här: REPL//AP-AP KN(och) AP → AP-AP eller så här REPL//AP-AP KN(- eller) AP → AP-AP. I det första av dessa två exemplen läggs en parentes med de KN som ska matchas av regeln och i det andra exemplet så innehåller en liknade parentes ord som inte ska matchas. På så vis kan reglerna göras mer specifika så att inte oönskade omskrivningar äger rum.

Regeln REPL//PP-PP NP → §P(S) är till för att plocka bort information om från vad för land något kommer ifrån eller ligger i. Den har också sina brister eftersom det inte är önskvärt att ta bort information om vart ett föremål ligger. Till exempel så ska inte meningen *Katten kommer ifrån kattlådan* göras om till *katten kommer*, då betydande information i meningen försvinner. Önskvärt är att kunna skriva regeln på ett mer

detaljerat vis så att den bara tar bort prepositionsfrasen om den föregås av ett adjektiv och där nominalfrasen innehåller ett egennamn snarare än ett annat substantiv.

Enligt Centrum för lättläst (2002) och Davidsson et al (2002) blir en text mer lättläst om färre bisatser används. På grund av detta har följande regler implementerats i CogFLUX:
REPL//NP-PM MID S MID → NP-PM Â§P(S)
REPL//NP-NN MID S MID → NP-NN Â§P(S)
Dessa regler är tänkta att endast ta bort bisatser som innehåller en adjektivfras och inget mer. För tillfället tas för många bisatser bort för önskat resultat och för att ge önskad effekt krävs det möjligheter till att bryta upp bisatsen mitt i regeln för att kolla upp så att bisatsen inte innehåller information som är relevent att behålla. En sådan regel skulle tillexempel kunna se ut så här:
REPL//NP-PM MID S-AP MID → NP-PM Â§P(S)
Bisatser som innehåller mer information till exempel om bisatsen består av fler än tre ord skulle vara bättre att skriva om så att det blir en ny mening.

Regeln REPL//NP-DT NN S → NP-NN S Â§P(PP) är så pass specialiserad att det inte hittats några tillfällen när den gör om fraser som egentligen inte bör göras om. Det förekommer säkert något fall där det sker en icke önskvärd syntaxförändring, men eftersom texterna som regeln är så pass långa tar det enorm tid att söka igenom varje mening.

6.3 Svårimplementerade regler

En del av de regler som Decker beskriver är svåra för ett datorprogram att hantera även om det är enkelt att beskriva på vilket vis frasen ska ändras. Alla regler som nu är implementerade i programkoden tar bort något från texten. Det är inte för att texter skulle bli mer lättlästa om de består av mindre text generellt. Det är snarare så att det för ett datorprogram är mycket enklare att plocka bort information än att lägga till information. Ett datorprogram kan bara arbeta med det som den får in. Att lägga till helt ny text som inte bara är omskrivningar av de ord som ges till programmet är mycket svårt att implementera. Alla omskrivningsregler som har fler och/eller andra enheter i den tänka omskrivningsfrasen är därför mycket svåra att få ett datorprogram att hantera. Ett exempel på en sådan regel som är svår att implementera är: np(ap+n[defness=indef]) → np(det+ap+n[defness=def]). I originaltexten så innehåller frasen en adjektivfras och

ett nomen men efter omskrivningsregeln så ska den nya texten innehålla förutom både en adjektivfras och ett nomen även en determinator. Det finns inte så många determinatorer i det svenska språket men tillräckligt många för att det inte ska gå att bara använda en enda för varje omskrivning.

6.4 Problem med Synonymhanteraren

Eftersom OVIX räknas ut beroende på hur många ord av varje ordklass texten innehåller så bör den inte ge någon skillnad i texten efter körningarna (Rybing & Smith 2009). Anledningen till att det blir en viss skillnad kan bero på att synonymhanteraren inte tar hänsyn till vilken ordklass ordet tillhör vilket innebär att den felaktigt kan byta ut vissa ord mot ett ord från en annan ordklass. Ett exempel är ordet *ett*, som både kan vara ett räkneord och artikel. Eftersom ordet *en* är en synonym till räkneordet *ett* men inte till partikeln ett så kan här ske ett felaktigt synonymutbyte. För att lösa problemet går det att använda sig av flera metoder. Ett sätt att lösa det på är att ha en ordklasstaggad synonymordbok där bara ord inom samma ordklass kan bytas ut. Ett annat mindre säkert tillvägagångssätt är att utnyttja folkets synonymordlistas, Synlex, egen gradering på hur likvärdiga synonymerna är. Enligt Kann (2003) sker inget synonymutbyte om värdet är för lågt, det vill säga under tre. Detta skulle troligen inte lösa problemet helt med utbyte av ord från olika ordklasser men däremot skulle det innebära att texten blir mer innehållsmässigt lik originalet. Ord kan vara dubbeltydiga även om de tillhör samma ordklasser. Ordet plan kan betyda både flygfarkost, ett område där sport utövas eller som en kortare variant av ordet planering. Alla dessa ordexempel är substantiv men beroende på sammanhanget så får de olika betydelser. Meningen *vi behöver en plan* får olika betydelser beroende på om den sägs av en grupp fotbollsspelare som vill ha någonstans att spela eller om det istället sägs av ett gäng banditer som tänker bryta sig in i en bank. Detta problem kan lösas bättre genom att ta hänsyn till den gradering som finns i folkets synonymordlista. Ytterligare förbättringar av synonymhanteringen kan göras genom att lägga till fler synonymer i listan.

6.5 Förbättring av synonymhanteraren

Synonymhanteraren hanterar just nu endast ord som har en synonym vars ordlängd är kortare än orginalordet. Detta ger förbättrade LIX-värden men inte nödvändigtvis en mer lättläst text. Ett exempel på ett ord som byts ut med synonymhanteraren är ”pojke” som

byts ut mot ”kis”. Även om kis är ett kortare ord känns det inte mer lättläst eftersom det inte är lika vanligt förekommande. Synonymhanteraren bör därför ta hänsyn till flera aspekter av ett ord än endast ordlängden. För att ett ord ska få bytas ut mot ett annat bär det bli mer lättläst utifrån fler kriterier än ordlängden. Vanligt förekommande ord är lättare att känna igen och på så vis även lättare att läsa. Ett sätt att få orden i texten att vara mer bekanta är att se till att texten innehåller så få unika ord som möjligt. Hittas ett ord som finns med i synlex så kan antalet identiska ord i texten räknas för att sedan jämföras med antalet förekomster av synonymkandidaten. På så vis går det att avgöra vilket ord som är vanligast i texten. Är Synonymförslaget mer vanligt förekommande än originalordet så kan originalordet bytas ut. Genom att ta hänsyn till detta får den nya texten mindre ordvarition och bör därför även ge ett förbättrat OVIX-värde. Ett annat sätt att byta ut ovanliga ord är att jämföra det mot en textdatabas till exempel Stockholm Umeå Corpus, SUC. Om ett ord är ovanligt och bör bytas ut så kan det göras genom att först jämföra ordet med en synonymkandidat för att se vilket ord som förekommer flest gånger i SUC. Det ord som finns med flest gånger kan betraktas som vanligast. Är originalordet mer vanligt förekommande än synonymkandidaten så får det ordet stå kvar och om synonymen är mer vanligt förekommande så byts ordet ut. Även om det är bra att införa fler vis att hantera synonymer så kan det innebära problem. Om ett kort ord har en synonymkandidat som är längre än men mer vanligt förekommande än originalordet så är det svårt att avgöra vilket av orden som ska användas. Ska ovanliga ord bytas ut mot längre och vanligare ord eller ska långa ord bytas ut mot korta men mer ovanliga ord?

Synlex innehåller en gradering för hur lika orden är. Denna gradering kan vara användbar för att avgöra om ett ord bör bytas ut eller ej eller för att värdera flera olika synonymkandidater. Synonymer är sällan ett helt perfekt substitut till originalordet. Olika ord används på olika vis även om det är synonymer. Meningen ”Kan jag anta att du hjälper till?” innehåller ordet anta som har flera synonymer men alla lämpar sig inte lika bra som substitut. Att byta ut ordet anta mot förmoda skulle inte göra någon direkt skillnad men att istället byta ut ordet anta mot gissa eller tro så känns inte meningen alls lika korrekt. I de fallen kan det vara bra att byta ut ordet anta mot ett ord som är både längre och mindre vanligt nämligen förmoda istället för det kortare och mer vanliga alternativet tro. Synonymerna till anta värderas som 4.7 för ordet förmoda och 3. 3 för ordet tro i synlex. Genom att bara byta ut ord som har ett högt likhetsvärde blir inte

risken att meningarna blir märkliga lika stor, men samtidigt så minskar antalet möjliga synonymutbyten markant om till exempel bara ord som värderas som 4.0 eller högre byts ut.

Synonymhanteraren klarar inte av olika böjningar av ord. Ordet byggnad kan bytas ut mot ordet hus men ordet byggnader kan inte bytas ut mot ordet hus. Detta är en mycket stor brist i synonymhanteraren eftersom väldigt många ord är böjda. För att förbättra synonymhanteraren så att även plural och bestämd form av ord kan bytas ut mot en motsvarande form krävs det att originalordets böjningsform analyseras. Den böjningsformen som originalordet kan appliceras på synonymen. I synlex står endast ordets grundform så för att kunna hitta en synonymkandidat måste även originalordets grundform eller lemma användas. Lemman kan jämföras mot synonym lexikonet för att hitta synonymkandidater. Hittas en synonym kan den bytas ut och skrivas om till samma böjningsform som originalordet ade.

Ett ytterligare påbyggnad av systemet med synonymer är att lägga till förklaringar av olika begrepp. En sådan hantering skulle kunna byta termer av ord som personer bekanta med domänen begriper till en utskriven förklaring av vad den termen står för. Något som är uppenbart för en person kan vara obegripligt för en annan. Ett exempel på ett utbyte som skulle kunna göras är att byta ut Stockholm mot Sveriges huvudstad.

Slutsats

Huvudresultatet visade att texter som förenklats med hjälp av SOR och synonymhanteraren får signifikant förbättrade värden då läsbarhetsmåtten LIX och nominalkvot används. Arbete som utförts tydliggör att det är fullt möjligt att skapa ett program som automatiskt skriver om texter så de blir mer lättlästa, enligt läsbarhetsmått. CogFlux har potential till att utvecklas vidare men upplevs som ett välfungerande system som gör texter tillgängliga till en större publik. För att utveckla programmet ytterligare skulle man till exempel kunna implementera fler regler och ta hänsyn till fler av de kriterier som gör en text lättläst. Som det ser ut idag går det att skriva om texter till mer lättläst svenska enligt läsbarhetsmåtten LIX, OVIX och Nominalkvot, vilket också var syftet med detta arbete.

Litteraturförteckning

Björnsson H. C. (1968) Läsbarhet. Bokförlaget Liber AB

Carlberger J. & Kann V. (1999) Implementing an efficient part-of-speech tagger *Nada, Numerical Analysis and Computing Science Royal Institute of Technology*

Centrum för lättläst (2002) Lättläst - vad är det? Hämtat 2011-04-25 från: http://www.lattlast.se/om-oss/lattlast---vad-ar-det

Dannélls D. (2010) Automatic generation and simplification of written documents *Göteborg University Department of Swedish language*

Davidsson J., Lönnborg T., Nyberg Å., Stymne S., Wahlberg K. & Ventura S (2002) SkrivLätt – en undersökning av möjligheterna att utveckla ett datoriserat hjälpmedel för framställning av lättlästa texter *Linköpings universitet Institutionen för datavetenskap*

Decker A. (2003) Towards automatic grammatical simplification of Swedish text *Stockholm University Department of Linguistics Computational Linguistics*

Freyoff G., Hess G., Kerr L., Menzel E., Tronbacke B. & Van der Veken K. (1998) Gör det enkelt skriv lätt *ILSMH European Association. Cascais: Cercia*

Hall (2006) MaltParser – An Architecture for Inductive Labeled Dependency Parsing *Växjö University School of Mathematics and Systems Engineering*

Heiman G. W. (2001) Understanding Research Methods and Statistics, Second Edition, Houghton Mifflin Company Boston New York

Jörgensen N. & Svensson J. (2004) Nusvensk gramatik Gleerups Utbildning AB

Kann V. (2004) Folkets användning av Lexin – en resurs *KTH Nada*

Kann V. & Rosell M. (2003) Free Construction of a Free Swedish Dictionary of Synonyms *KTH Nada*

Korhonen (2010) Läsbarhet som bedömningsmetod hos uppsatser skrivna av finska universitetsstuderande *Jyväskylä universitet Institutionen för språk Svenska språket*

Lundberg I. & Reichenberg M. (2008) Vad är lättläst? *Specialpedagogiska skolmyndigheten*

Mühlenbock K. & Johansson Kokkinakis S. (2009) LIX 68 revisited An extended readability measure *Department of Swedish Gothenburg university*

Nivre J. & Hall J. (2009) A Quick Guide to MaltParser Optimization

Petersen S. E. & Ostendorf M. (2007) Text simplification for language learners: a corpus analysis *Dept. of Computer Science, Dept. of Electrical Engineering,University of Washington*

Regeringskansliet (2009) Språk för alla information om Sveriges nya språklag Hämtat 2011-04-25 från: http://www.regeringen.se/content/1/c6/12/92/67/c7740716.pdf

Rybing J. & Smith C. (2009) CogFLUX – Grunden till ett automatiskt textförenklingssystem för svenska *Linköpings universitet*

Språklag 2009:600 (2009) Hämtat 2011-04-25 från Svensk författningssamling: http://www.riksdagen.se/webbnav/index.aspx?nid=3911&bet=2009:600

Wilhelmsson T. (2007) Texters läsbarhet ur ett andraspråksperspektiv – En jämförande studie mellan två versioner av ett informationsmaterial som riktas till invandrade *Göteborgs Universitet Institutionen för svenska språket Svenska som andraspråk*

Deckerregler

np(n1+pp(p+np(n2))) → np2

np1+vp+pp(p+np2(n)) → np1(n+som+aux+vpass))+vp

np(n1-s-n2) → np(n1)

np(n1[number=sg]-n2) → np(n2+pp(p+np(n1[number=pl])))

np(det+n+X) → np(n+X)

np(det+ap+n[defness=def]) → np(ap+n[defness=indef])

np(det+ap+n) → np(n)

np(ap+n) → np(n)

np(ap+n[defness=indef]) → np(det+ap+n[defness=def])

np(ap+nn[defness=indef]) → np(nn[defness=def]+pp(p+np(pn)))

np(det+n:poss[defness=indef]+n[defness=indef])→

np(n[defness=def+pp(p+np(n[defness=def]))

np(pn:poss+n) → np(n)

np(n+pp) → np(n)

np(n1+pp(p+np(n2))) → np(n2+(som+pron+v))

np(n+subj+X) → np(n)

ap(adj) → Ø

ap(adj1+ik+adj2) → ap(adj1)

ap(adj1+konj+adj2) → ap(adj2)

ap(adv+adj) → ap(adj)

pp(p+np) → Ø

pp(p1+np(n+pp(p2+np1)) → som+v+pp(p1+np1))

pp(p+np) → np+(som+v+pp(p+np)))

pp(p+np1) → np1+v+np

konj+{main clause} → new sentence

formsubj+vkop+np(n:poss+subclause) → np+aux+vp(v+subclause)

Utdrag ur Folkets synonymlexikon Synlex

<syn level="4.3"><w1>abakus</w1><w2>kulram</w2></syn>
<syn level="5.0"><w1>abbreviera</w1><w2>förkorta</w2></syn>
<syn level="4.0"><w1>abdikera</w1><w2>avgå</w2></syn>
<syn level="5.0"><w1>abdikera</w1><w2>avsäga sig tronen</w2></syn>
<syn level="3.2"><w1>abnorm</w1><w2>missbildad</w2></syn>
<syn level="3.2"><w1>abnorm</w1><w2>onaturlig</w2></syn>
<syn level="3.2"><w1>abnorm</w1><w2>onormal</w2></syn>
<syn level="3.4"><w1>abnormitet</w1><w2>avvikelse</w2></syn>
<syn level="3.6"><w1>abonnent</w1><w2>prenumerant</w2></syn>
<syn level="4.0"><w1>abonnera</w1><w2>boka</w2></syn>
<syn level="4.0"><w1>abonnera</w1><w2>prenumerera</w2></syn>
<syn level="4.0"><w1>abrupt</w1><w2>plötslig</w2></syn>
<syn level="4.0"><w1>abrupt</w1><w2>plötsligt</w2></syn>
<syn level="5.0"><w1>abrupt</w1><w2>tvärt</w2></syn>
<syn level="3.0"><w1>absolut</w1><w2>bestämt</w2></syn>
<syn level="4.0"><w1>absolut</w1><w2>definitiv</w2></syn>
<syn level="4.0"><w1>absolut</w1><w2>definitivt</w2></syn>
<syn level="3.6"><w1>absolut</w1><w2>exakt</w2></syn>
<syn level="3.1"><w1>absolut</w1><w2>fullkomlig</w2></syn>
<syn level="3.2"><w1>absolut</w1><w2>fullkomligt</w2></syn>
<syn level="3.0"><w1>absolut</w1><w2>helt</w2></syn>
<syn level="3.6"><w1>absolut</w1><w2>otvetydig</w2></syn>
<syn level="4.0"><w1>absolut</w1><w2>otvivelaktigt</w2></syn>
<syn level="3.6"><w1>absolut</w1><w2>ovillkorlig</w2></syn>
<syn level="3.8"><w1>absolut</w1><w2>ovillkorligen</w2></syn>
<syn level="3.0"><w1>absolut</w1><w2>precis</w2></syn>
<syn level="3.0"><w1>absolut</w1><w2>prompt</w2></syn>
<syn level="4.1"><w1>absolut</w1><w2>självklart</w2></syn>
<syn level="4.0"><w1>absolut</w1><w2>säkert</w2></syn>
<syn level="3.4"><w1>absolut</w1><w2>total</w2></syn>